Résolution d'un système linéaire par méthodes itératives

1) Principe de base

Soit une fonction réelle f à variable réelle, dérivable sur un intervalle I et telle que :

$$\forall\, x \in I : \; |f'(x)| \leq k < 1$$

et soit la suite définie par :

$$\begin{cases} x_{p+1} = f(x_p) \\ x_0 \in I \end{cases}$$

alors

a) **l'équation $f(x) = x$ admet une unique solution a sur I**
b) **la suite (x_p) converge vers a**

Preuve : Nous allons montrer que la suite (x_p) vérifie le critère de Cauchy pour la convergence. Nous avons en effet pour tout entier naturel p :

$$x_{p+2} - x_{p+1} = f(x_{p+1}) - f(x_p)$$

et d'après le théorème des accroissements finis :

$$\exists\, c_p \in \left]x_p, x_{p+1}\right[\left(ou \left]x_{p+1}, x_p\right[\right) \; : \qquad f(x_{p+1}) - f(x_p) = (x_{p+1} - x_p)\, f'(c_p)$$

soit :

$$|x_{p+2} - x_{p+1}| = |x_{p+1} - x_p|\, |f'(c_p)|$$

D'où :

$$|x_{p+2} - x_{p+1}| \leq k|x_{p+1} - x_p|$$

On en déduit par une récurrence simple :

$$|x_{p+1} - x_p| \leq k^p\, |x_1 - x_0|$$

Puis $\forall\, (p, n) \in \mathbb{N} \times \mathbb{N}^*$:

$$|x_{p+n} - x_p| = |x_{p+n} - x_{p+n-1} + x_{p+n-1} - x_{p+n-2} + \cdots + x_{p+1} - x_p|$$

$$|x_{p+n} - x_p| \leq \; |x_{p+n} - x_{p+n-1}| + |x_{p+n-1} - x_{p+n-2}| + \cdots + |x_{p+1} - x_p|$$

$$|x_{p+n} - x_p| \leq (k^{p+n-1} + k^{p+n-2} + \cdots + k^p)|x_1 - x_0|$$

$$|x_{p+n} - x_p| \leq k^p(k^{n-1} + k^{n-2} + \cdots + 1)|x_1 - x_0|$$

$$|x_{p+n} - x_p| \leq k^p \frac{1 - k^n}{1 - k}|x_1 - x_0|$$

$$|x_{p+n} - x_p| \leq \frac{k^p}{1 - k}|x_1 - x_0|$$

Soit alors $\varepsilon > 0$. On a :

$$\exists\, n_0 \in \mathbb{N} : p > n_0 \Rightarrow \frac{k^p}{1 - k}|x_1 - x_0| < \varepsilon$$

Soit :

$$p > n_0\,, n \in \mathbb{N}^* \Rightarrow |x_{p+n} - x_p| < \varepsilon$$

donc la suite (x_p) est de Cauchy donc converge vers une limite a de I

Par passage à la limite dans la relation de récurrence, a vérifie :

$$f(a) = a$$

La limité d'une suite étant unique, l'équation $f(x) = x$ a donc une unique solution sur I

2) Version vectorielle en dimension finie

Soit $\mathbb{E}$ l'espace vectoriel des matrices colonnes à n lignes à coefficients dans un corps $\mathbb{K}$ ($\mathbb{R}$ou $\mathbb{C}$). Soit $B \in \mathbb{E}$, M une matrice carrée d'ordre n et f l'application affine de $\mathbb{E}$ dans $\mathbb{E}$ définie par :

$$\boldsymbol{f(X) = M\,X + B}$$

Soit V l'ensemble fini des valeurs propres complexes de M.On désigne par rayon spectral de M la quantité :

$$\boldsymbol{\rho(M) = sup_{\lambda \in V}|\lambda|}$$

Soit (X_p) une suite d'éléments de $\mathbb{E}$ alors :

$$\left(\forall\, \boldsymbol{Z} \in \mathbb{E} : \begin{cases} \boldsymbol{X_{p+1} = f(X_p)} \\ \boldsymbol{(X_p)\ \text{converge}} \\ \boldsymbol{X_0 = Z} \end{cases}\right) \Leftrightarrow \boldsymbol{\rho(M) < 1}$$

Preuve :

Rappelons d'abord que toute matrice complexe M est trigonalisable dans $\mathbb{C}$, plus précisément, il existe une matrice T triangulaire supérieure dont la diagonale $(\lambda_1, \lambda_2, \dots, \lambda_n)$ est formée des valeurs propres complexes de M et une matrice P inversible telles que :

$$P^{-1}\, M\, P = T$$

$(\Rightarrow)$

Supposons que (X_p) converge vers une limite L, alors la suite extraite (X_{p+1}) converge vers L et la suite $\left(f(X_p)\right)$ vers $f(L)$. Donc L vérifie :

$$L = f(L) = M\, L + B$$

La suite $(Y_p = X_p - L)$ tend 0 et vérifie la relation de récurrence :

$$Y_{p+1} = M\, Y_p$$

D'où on déduit par récurrence la forme explicite :

$$Y_p = M^p\, Y_0 = M^p\, (Z - L)$$

Le fait que (Y_p) tende vers 0 quelque soit la valeur de Z fait que chaque coefficient de M^p tend vers 0. En prenant en effet par exemple $Z - L$ égale à une colonne avec un 1 sur la première ligne et des 0 pour les suivantes, on obtient pour $M^p\, (Z - L)$ la première colonne de M^p. Or :

$$P^{-1}\, M^p\;\, P = T^p$$

où T^p est une matrice triangulaire supérieure dont la diagonale est $({\lambda_1}^p, {\lambda_2}^p, \dots, {\lambda_n}^p)$

Le fait que les coefficients de M^p tendent vers 0 implique que ceux de T^p tendent également vers 0, donc en particulier les éléments diagonaux, d'où :

$$\forall\, i \in \{1, \dots, n\} :\ |\lambda_i| < 1$$

Donc :

$$\rho(M) < 1$$

$(\Leftarrow)$

Supposons : $\rho(M) < 1$ et soit $Z\ \in \mathbb{E}$ et la suite (X_p) définie par :

$$\begin{cases} X_{p+1} = f(X_p) \\ \quad X_0 = Z \end{cases}$$

Montrons que $f(X)$ admet un unique point fixe en résolvant :

$$X = M\,X + B$$

$$\Leftrightarrow (I_n - M)\,X = B$$

Or :

$$P^{-1}\,(I_n - M)\,P = I_n - T$$

$I_n - T$ est une matrice triangulaire supérieure avec des coefficients diagonaux qui sont formés par les $1 - \lambda_i$ donc non nuls. Elle est donc inversible. On en déduit que $I_n - M$ l'est aussi. Ainsi :

$$(I_n - M)\,X = B$$

$$\Leftrightarrow\ X = (I_n - M)^{-1}\,B$$

Donc $f(X)$ admet un unique point fixe que nous noterons L. La suite $\left(Y_p = X_p - L\right)$ vérifie la relation de récurrence

$$Y_{p+1} = M\,Y_p$$

et a donc pour expression explicite :

$$Y_p = M^p\,Y_0 = M^p\,(Z - L)$$

Reste à montrer que les coefficients de M^p tendent tous vers 0 pour montrer que Y_p tend vers 0 et donc X_p vers L. Pour cela, il faut un peu plus que le caractère trigonalisable de M et invoquer la décomposition de Jordan dans laquelle la matrice T est formée de blocs diagonaux carrés d'ordre $k \leq n$ de la forme :

$$J_k = \begin{pmatrix} \lambda_i & 1 & 0 & 0 \\ 0 & \lambda_i & \ddots & \ddots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & \lambda_i \end{pmatrix} = \lambda_i\,I_k + N_k$$

Où N_k est une matrice nilpotente d'indice k.

Il en résulte que T^p est une matrice formée des blocs diagonaux ${J_k}^p$. Or d'après la formule du binôme de Newton (utilisable car $\lambda_k\,I_k$ et N_k commutent) , on a pour $p > n$:

$${J_k}^p = (N_k + \lambda_i\,I_k)^p = \sum_{q=0}^{p} \binom{p}{q}\,{N_i}^q(\lambda_i\,I_k)^{p-q} = \sum_{q=0}^{n} \binom{p}{q}\,{\lambda_i}^{p-q}{N_i}^q$$

Or : $|\lambda_i| \leq \rho(M) < 1$ donc :

$$\lim_{p\to+\infty} {\lambda_i}^{p-q} = 0$$

On en déduit que les coefficients de ${J_k}^p$ tendent vers 0 donc ceux de T^p

3) <u>Méthode de Gauss-Seidel</u>

Soit un système linéaire d'ordre n écrit sous forme matricielle :

$$A\,X = Y$$

où Y est un vecteur-colonne d'ordre n donné et A une matrice carrée d'ordre n. Décomposons A en une somme de trois matrices :

$$A = L + D + U$$

L (comme Lower en anglais) étant la matrice triangulaire inférieure formée des éléments situés strictement sous la diagonale de A, D la matrice diagonale formée des éléments diagonaux de A et U (comme Upper en anglais) la matrice formée des éléments situés strictement au dessus de la diagonale de A.

Le système peut alors s'écrire ainsi :

$$(L + D)\,X = -U\,X + Y$$

Or $L + D$ est une matrice triangulaire inférieure qui est inversible si et seulement si les éléments diagonaux de D sont tous non nuls. Le système devient alors :

$$X = -\,(L + D)^{-1}U\,X + (L + D)^{-1}Y$$

Posons :

$$M = -\,(L + D)^{-1}U$$

$$B = (L + D)^{-1}Y$$

Alors le système s'écrit :

$$X = M\,X + B$$

La méthode de Gauss Seidel consiste alors à le résoudre de façon itérative en introduisant la suite définie par :

$$\begin{cases} X_{p+1} = f\left(X_p\right) \\ X_0 = Z \end{cases}$$

Z étant un vecteur colonne d'ordre n quelconque.

Une condition nécessaire de convergence de la suite vers la solution du système est, comme nous l'avons vu précédemment :

$$\rho(M) < 1$$

Notons que l'itération s'écrit :

$$X_{p+1} = D^{-1}\left(Y - U\,X_p - L\,X_{p+1}\right)$$

Autrement dit, les composantes de X_{p+1} s'obtiennent comme suit :

$$\begin{cases} X_{p+1}[1] = \dfrac{1}{a_{11}}\left(Y[1] - \displaystyle\sum_{i=2}^{n} a_{1i}\,X_p[i]\right) \\ X_{p+1}[2] = \dfrac{1}{a_{22}}\left(Y[2] - \displaystyle\sum_{i=3}^{n} a_{2i}\,X_p[i] - a_{21}\,X_{p+1}[1]\right) \\ \vdots \\ X_{p+1}[q] = \dfrac{1}{a_{qq}}\left(Y[q] - \displaystyle\sum_{i=q+1}^{n} a_{qi}\,X_p[i] - \sum_{i=1}^{q-1} a_{qi}\,X_{p+1}[i]\right) \\ \vdots \\ X_{p+1}[n] = \dfrac{1}{a_{nn}}\left(Y[n] - \displaystyle\sum_{i=1}^{n-1} a_{ni}\,X_{p+1}[i]\right) \end{cases}$$

D'un point de vue algorithmique, il convient de réserver n^2 places en mémoire pour le stockage des coefficients de la matrice A, n places pour le stockage des coefficients du second membre Y et n places pour le stockage de la solution du système X soit au total :

$n^2 + 2\,n$ coefficients réels ou complexes à stocker

L'algorithme s'écrit alors simplement sous la forme :

Initialisation de $X, T = 1$

Boucle tant que $T > précision\ choisie$

Boucle pour q allant de 1 à n :

$$T = 0$$

$$Z = \frac{1}{a_{qq}}\left(Y[q] - \sum_{i=1, i\neq q}^{n} a_{qi}\,X[i]\right)$$

$$T = Max(|X[q] - Z|, T)$$

$$X[q] = Z$$

Fin de la boucle sur q

Fin Tant que

Faisons le décompte des opérations pour chaque itération. Pour chacune des valeurs de q il y a à exécuter :

- $n-1$ multiplications
- n additions
- 1 division

Ile nombre d'opérations pour chaque itération est :

$$n\,(n-1+n+1)=2\,n^2$$

<u>Exemple avec un système d'ordre 2</u>

Soit le système suivant à résoudre :

$$\begin{cases} 2\,x+y=4 \\ x+5\,y=11 \end{cases}$$

La solution est le couple $(1;2)$ mais nous allons la retrouver avec la méthode de Gauss Seidel, qui évidemment ne serait pas adaptée pour un système de cet ordre pouvant se résoudre directement.

Le processus de récurrence se traduit par :

$$\begin{cases} 2\,x_{p+1}+y_p=4 \\ x_{p+1}+5\,y_{p+1}=11 \end{cases}$$

Il conduit à :

$$\begin{cases} x_{p+1}=\dfrac{1}{2}\left(4-y_p\right) \\ y_{p+1}=\dfrac{1}{5}\left(11-x_{p+1}\right) \end{cases}$$

Voyons ce que donnent les 4 premières itérations en partant d'un vecteur initial :

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} x_1 = \frac{1}{2}(4 - y_0) = 2 \\ y_1 = \frac{1}{5}(11 - x_1) = \frac{9}{5} = 1{,}8 \end{cases}$$

$$\begin{cases} x_2 = \frac{1}{2}(4 - y_1) = 1{,}1 \\ y_2 = \frac{1}{5}(11 - x_2) = \frac{9{,}9}{5} = 1{,}98 \end{cases}$$

$$\begin{cases} x_3 = \frac{1}{2}(4 - y_2) = 1{,}01 \\ y_3 = \frac{1}{5}(11 - x_3) = 1{,}998 \end{cases}$$

$$\begin{cases} x_4 = \frac{1}{2}(4 - y_3) = 1{,}001 \\ y_4 = \frac{1}{5}(11 - x_4) = 1{,}9998 \end{cases}$$

Nous constatons qu'après 4 itérations nous aboutissons à la valeur de la solution à 0,001 près. Vérifions que nous sommes bien dans un cas de convergence. Nous avons :

$$M = -(L + D)^{-1} U = -\begin{pmatrix} 2 & 0 \\ 1 & 5 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

$$= -\frac{1}{10}\begin{pmatrix} 5 & 0 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -0{,}5 \\ 0 & 0{,}1 \end{pmatrix}$$

Les valeurs propres de M sont donc : 0 et 0,1 donc :

$$\rho(M) = 0{,}1 < 1$$

La convergence est donc assurée mais nous verrons plus loin une condition plus facile à observer.

<u>Exemple avec un système d'ordre 3</u>

Soit le système suivant à résoudre :

$$\begin{cases} x + 2\,y - z = 6 \\ 2\,x + y + z = 9 \\ x - y - z = 0 \end{cases}$$

La solution est le triplet $(3; 2; 1)$ voyons ce que donne la méthode de Gauss Seidel

Le processus de récurrence se traduit par :

$$\begin{cases} x_{p+1} + 2\,y_p - z_p = 6 \\ 2\,x_{p+1} + y_{p+1} + z_p = 9 \\ x_{p+1} - y_{p+1} - z_{p+1} = 0 \end{cases}$$

Il conduit à :

$$\begin{cases} x_{p+1} = 6 - 2\,y_p + z_p \\ y_{p+1} = 9 - 2\,x_{p+1} - z_p \\ z_{p+1} = x_{p+1} - y_{p+1} \end{cases}$$

Partant de :

$$\begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

voyons les 4 premières itérations (un tableur permet d'effectuer aisément les calculs) :

$$\begin{cases} x_1 = 6 - 2\,y_0 + z_0 = 6 \\ y_1 = 9 - 2\,x_1 - z_0 = -3 \\ z_1 = x_1 - y_1 = 9 \end{cases}$$

$$\begin{cases} x_2 = 6 - 2\,y_1 + z_1 = 21 \\ y_2 = 9 - 2\,x_2 - z_1 = -24 \\ z_2 = x_2 - y_2 = 45 \end{cases}$$

$$\begin{cases} x_3 = 6 - 2\,y_2 + z_2 = 99 \\ y_3 = 9 - 2\,x_3 - z_2 = -144 \\ z_3 = x_3 - y_3 = 243 \end{cases}$$

$$\begin{cases} x_4 = 6 - 2\,y_3 + z_3 = 537 \\ y_4 = 9 - 2\,x_4 - z_3 = -822 \\ z_4 = x_4 - y_4 = 1359 \end{cases}$$

Il semble que la suite ne soit pas convergente. Essayons d'en déterminer la cause. Nous avons :

$$M = -\,(L+D)^{-1}U = -\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & -1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 2 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$$= -\begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & -1 & -1 \end{pmatrix} \begin{pmatrix} 0 & 2 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -2 & 1 \\ 0 & 4 & -3 \\ 0 & -6 & 4 \end{pmatrix}$$

Le polynôme caractéristique de M est :

$$P_M(X) = \begin{vmatrix} -X & -2 & 1 \\ 0 & (4-X) & -3 \\ 0 & -6 & (4-X) \end{vmatrix} = -X\,((4-X)^2 - 18)$$

$$P_M(X) = -X\left(4 - 3\sqrt{2} - X\right)\left(4 + 3\sqrt{2} - X\right)$$

Les valeurs propres de M sont donc : $4 - 3\sqrt{2}\,, 0, 4 + 3\sqrt{2}$

Le rayon spectral de M est donc :

$$\rho(M) = 4 + 3\sqrt{2} > 1$$

La convergence n'est donc pas assurée.

<u>Condition suffisante de convergence de la méthode de Gauss-Seidel</u>

Le calcul du rayon spectral de M n'étant pas aisé pour des matrices d'ordre important, il est utile de disposer d'un critère de convergence suffisant facile à évaluer. Le théorème d'Hadamard et le théorème de Gershgorin vont préciser ces questions.

<u>Théorème d'Hadamard</u> :

Si M est une matrice carrée d'ordre n non inversible (on dit singulière) alors il existe un élément diagonal dont le module est inférieur ou égal à la somme des modules des autres termes de sa ligne et il existe un élément diagonal dont le module est inférieur ou égal à la somme des modules des autres termes de sa colonne.

<u>Preuve</u> :

Soit X un vecteur colonne non nul tel que : $M\,X = 0$

Soit $i \in \{1, \dots, n\}$ tel que : $X[i] = max\{X[1], X[2], \dots, X[n]\}$

Alors :

$$m_{ii}\,X[i] = -\sum_{j=1, j\neq i}^{n} m_{ij}\,X[j]$$

Donc :

$$|m_{ii}|\,|X[i]| \le \sum_{j=1, j\neq i}^{n} |m_{ij}|\,|X[j]| \le |X[i]| \sum_{j=1, j\neq i}^{n} |m_{ij}|$$

d'où :

$$|m_{ii}| \leq \sum_{j=1, j\neq i}^{n} |m_{ij}|$$

De plus si M est singulière alors sa transposée l'est donc il existe $i \in \{1, \dots, n\}$ tel que :

$$|m_{ii}| \leq \sum_{j=1, j\neq i}^{n} |m_{ji}|$$

Théorème de Gershgorin :

C'est une conséquence du précédent.

Si M est une matrice carrée d'ordre n et λ une valeur propre de M alors $M - \lambda\, I_n$ est singulière donc il existe $i \in \{1, \dots, n\}$ tel que :

$$|m_{ii} - \lambda| \leq \sum_{j=1, j\neq i}^{n} |m_{ij}|$$

et il existe $i \in \{1, \dots, n\}$ tel que :

$$|m_{ii} - \lambda| \leq \sum_{j=1, j\neq i}^{n} |m_{ji}|$$

Ce théorème permet de localiser les valeurs propres dans des cercles du plan complexe ce qui peut être utile pour vérifier que le rayon spectral est strictement inférieur à 1.

Un critère simple conséquence du théorème d'Hadamard :

Reprenons les notations de la méthode de Gauss Seidel. Soit λ une valeur propre de M alors il existe $X \neq 0$ tel que :

$$M\, X = \lambda\, X$$

Soit :

$$-\,(L + D)^{-1} U\, X = \lambda\, X$$

$$U\, X = \lambda\, (L + D) X$$

$$(\lambda\, (L + D) - U)\, X = 0$$

Donc la matrice $\lambda\, (L + D) - U$ est singulière. On en déduit d'après le théorème de d'Hadamard l'existence d'un indice $i \in \{1, \dots, n\}$ tel que :

$$|\lambda\, a_{ii}| \leq \sum_{j=1}^{i-1} |\lambda\, a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|$$

Soit :

$$|\lambda|\,|a_{ii}| \leq |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|$$

Supposons alors que l'on ait :

$$\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| < |a_{ii}|$$

Alors :

$$|\lambda| \left(\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| \right) < |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|$$

D'où :

$$|\lambda| \sum_{j=i+1}^{n} |a_{ij}| < \sum_{j=i+1}^{n} |a_{ij}|$$

Donc :

$$|\lambda| < 1$$

Ainsi :

$$\rho(M) < 1$$

La méthode de Gauss Seidel converge alors

Une condition suffisante de convergence de la méthode de Gauss Seidel est donc :

$$\forall\, \boldsymbol{i} \in \{\mathbf{1}, \dots, \boldsymbol{n}\} : \sum_{j=1}^{i-1} |\boldsymbol{a_{ij}}| + \sum_{j=i+1}^{n} |\boldsymbol{a_{ij}}| < |\boldsymbol{a_{ii}}|$$

Ou bien

$$\forall\, \boldsymbol{i} \in \{\mathbf{1}, \dots, \boldsymbol{n}\} : \sum_{j=1}^{i-1} |\boldsymbol{a_{ji}}| + \sum_{j=i+1}^{n} |\boldsymbol{a_{ji}}| < |\boldsymbol{a_{ii}}|$$

Autrement dit que chaque élément diagonal soit de module strictement supérieur à la somme des modules des termes de sa ligne ou que chaque élément diagonal soit de module strictement supérieur à la somme des modules des termes de sa colonne.

Revenons sur le second exemple de matrice :

$$A = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 1 \\ 1 & -1 & -1 \end{pmatrix}$$

Le premier élément diagonal vaut 1 et la somme des modules des autres éléments de sa ligne vaut 3 et la somme des modules des autres éléments de sa colonne vaut 3 également. Le critère n'était donc pas vérifié.

En revanche pour le premier exemple de matrice :

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 5 \end{pmatrix}$$

le critère était vérifié ($1 < 2$ et $1 < 5$) , la convergence était donc assurée par la simple observation de ce critère.